IMPLEMENTASI TEKNIK DATA MINING UNTUK MEMPREDIKSI TINGKAT KELULUSAN MAHASISWA PADA UNIVERSITAS BINA DARMA PALEMBANG

Andri[1)], **Yesi Novaria Kunang** [2)], **Sri Murniati** [3)]
[1,2,3)]Jurusan Sistem Informasi Universitas Bina Darma Palembang
Jl. Jend. A. Yani No.12 Palembang Telp (0711)-515879
e-mail : andri@mail.binadarma.ac.id

Abstrak
Kegiatan evaluasi, perencanaan, dan pengambilan keputusan akan dapat dilakukan dengan lebih baik jika sebuah organisasi memiliki informasi yang lengkap, cepat, tepat, dan akurat. Informasi yang dibutuhkan dapat diekstrak dari data operasional yang tersimpan dalam database yang terintegrasi. Penelitian ini mengkaji ektraksi data operasional ke dalam sebuah data warehouse untuk kemudian dilanjutkan dengan kegiatan analisis data menggunakan teknik data mining. Data Mining merupakan proses analisis data menggunakan perangkat lunak untuk menemukan pola atau aturan tertentu dari sejumlah data dalam jumlah besar yang diharapkan dapat menemukan pengetahuan guna mendukung keputusan. Penelitian ini memanfaatkan data Mahasiswa dan data IPK, untuk menentukan karakteristik mahasiswa yang digunakan untuk prediksi kelulusan. Dalam penelitian ini teknik data mining yang digunakan yaitu Classification dengan menerapkan metode Decision Tree dan algoritma J48 untuk membantu menemukan karakteristik atau variabel yang mempengaruhi tingkat kelulusan mahasiswa pada jurusan sistem informasi Universitas Bina Darma Palembang, sehingga untuk selanjutnya dapat digunakan dalam memprediksi tingkat kelulusan mahasiswa yang akan datang. Tools yang digunakan untuk proses analisis data mining dalam penelitian ini menggunakan Weka 3.6.8. Dari hasil analisis yang telah dilakukan yang menggunakan data Mahasiswa dan IPK sebagai sampel dihasilkan keputusan bahwa variabel tempat lahir memiliki nilai Gain tertinggi sehingga atribut ini menjadi root dalam Decision Tree, kesimpulan akhir didapat bahwa variabel tempat lahir, pekerjaan orang tua, asal sekolah dan jenis kelamin adalah variabel yang menentukan tingkat kelulusan mahasiswa pada jurusan Sistem Informasi Universitas Binadarma Palembang.

***Kata Kunci** : Data Warehouse, Data Mining, Classification, Decision Tree, Algoritma J48*

1. PENDAHULUAN

1.1 Latar Belakang

Informasi yang akurat sangat dibutuhkan dalam kehidupan sehari-hari, informasi akan menjadi suatu elemen penting dalam perkembangan masyarakat saat ini dan waktu mendatang.Pemanfaatan data yang ada di dalam sistem informasi untuk menunjang kegiatan pengambilan keputusan tidak cukup hanya mengandalkan data operasional saja, tetapi diperlukan suatu analisis data untuk menggali potensi-potensi informasi yang ada. Para pengambil keputusan berusaha untuk memanfaatkan gudang data yang sudah dimiliki dalam mengambil keputusan, hal ini mendorong munculnya cabang ilmu baru untuk mengatasi masalah penggalian informasi atau pola yang penting dan menarik dari data jumlah besar, yang disebut dengan data mining. Penggunaan teknik data mining diharapkan dapat memberikan pengetahuan-pengetahuan yang sebelumnya tersembunyi di dalam gudang data (*data warehouse*), sehingga menjadi informasi yang berharga.

Perguruan tinggi saat ini dituntut untuk memiliki keunggulam bersaing dengan memanfaatkan semua sumber daya yang dimiliki. Sistem informasi digunakan untuk mendapatkan, mengolah dan menyebarkan informasi serta menunjang kegiatan operasional sehari-hari sekaligus menunjang kegiatan pengambilan keputusan strategis. Pemanfaatan data kelulusan pada gudang data saat ini belum dimanfaatkan secara maksimal dan efisien, sehingga tingkat kelulusan mahasiswanya belum sepenuhnya dilihat dengan mudah dan cepat. Untuk melihat dan dapat mengetahui prediksi tingkat kelulusan mahasiswa tersebut, maka dapat memaksimalkan dan memanfaatkan data-data yang menumpuk di gudang data khususnya data kelulusan. Dengan memanfaatkan teknik data mining, peneliti mencoba untuk menggali dan mendapatkan informasi-informasi yang berguna untuk mendukung pengambilan keputusan manajemen terutama yang berkaitan dengan prediksi tingkat kelulusan mahasiswa.

Berdasarkan latar belakang yang telah di uraikan diatas, maka permasalahan yang dapat dirumuskan dalam penelitian ini adalah "Bagaimana mengimplementasi teknik data mining untuk memprediksi tingkat kelulusan mahasiswa pada Universitas Bina Darma Palembang?". Tujuan dari penelitian ini adalah untuk mengetahui variabel-variabel yang mempengaruhi tingkat kelulusan mahasiswa jurusan Sistem Informasi Fakultas Ilmu Komputer Universitas Binadarma Palembang.

2. TINJAUAN PUSTAKA

Beberapa penelitian yang menjadi referensi dalam penelitian yang penulis lakukan meliputi, penelitian yang dilakukan oleh Ernawati yang berjudul "Prediksi Status Keaktifan Studi Mahasiswa dengan Algoritam C5.0 dan K-Nearest Neighbor", penelitian ini bertujuan menerapkan dan melakukan analisis algoritma C5.0 dan K-Nearest Neighbor pada dataset akademik untuk melihat karakteristik mahasiswa yang aktif, penelitian berikutnya adalah penelitian yang dilakukan oleh Raditya yang berjudul"Implementasi *Data Mining Classification* untuk mencari pola prediksi Hujan dengan menggunakan Algoritma C4.5. Penelitian ini mengunakan bahasa pemrograman *Java* serta DBMS *MySQL* untuk membangun aplikasinya. Akurasi pola prediksi yang didapat mampu mencapai 79%. Akurasi tersebut dihasilkan dari uji coba dengan mengunakan data cuaca tahun 2007 sebagai data training nya serta data cuaca tahun 2008 dan 2009 sebagai data testingnya. Penelitian berikutnya yang menjadi acuan adalah penelitian yang dilakukan oleh Azimah dan Sucahyo dengan judul "Penggunaan *Data Warehouse* dan *Data Mining* untuk Data Akademik", Tujuan penelitian ini adalah untuk mengetahui Kegiatan evaluasi, perencanaan, dan pengambilan keputusan akan dapat dilakukan dengan lebih baik jika sebuah organisasi memiliki informasi yang lengkap, cepat, tepat, dan akurat.

2.1 *Data Warehouse*

Data warehouse adalah sistem yang mengambil dan menggabungkan data secara periodik dari sistem sumber data ke penyimpanan data bentuk dimensional atau normal (Rainardi, 2008).
Menurut (Han, 2006) *Data Warehouse* adalah penyimpanan data tetap sebagai implementasi fisik dari pendukung keputusan model data. *Data warehouse* juga biasanya dilihat sebagai arsitektur, pembangunan dan penyatuan data dari bermacam-macam sumber data yang berbeda untuk mendukung struktur dan atau *query* tertentu, laporan analisis dan pembuatan keputusan. Data warehouse merupakan suatu lingkungan dimana *user* bisa menemukan suatu informasi yang strategis atas kumpulan data-data yang dimiliki, yang bersifat *integrated*, *subject-oriented*, *nonvolatile*, dan *time-variant*.

2.2 *Data Mining*

Data mining merupakan bidang dari beberapa bidang keilmuan yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistik, database, dan visualisasi untuk penanganan permasalahan pengambilan informasi dari database yang besar (Larose, 2005).
Secara garis besar data mining dapat dikelompokkan menjadi 2 kategori utama, yaitu (Tan et al, 2005) :

a) *Descriptive mining*, yaitu proses untuk menemukan karakteristik penting dari data dalam suatu basis data. Teknik data mining yang termasuk dalam *descriptive mining* adalah *clustering*, *association*, dan *sequential mining*.
b) *Predictive*, yaitu proses untuk menemukan pola dari data dengan menggunakan beberapa variabel lain di masa depan. Salah satu teknik yang terdapat dalam *predictive mining* adalah klasifikasi.

2.3 *Classification*

Classification adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui (Han dan Kamber, 2006). Klasifikasi merupakan fungsi pembelajaran yang memetakan (mengklasifikasi) sebuah unsur (*item*) data ke dalam salah satu dari beberapa kelas yang sudah didefinisikan. Data *input* untuk klasifikasi adalah koleksi dari *record*. Setiap record dikenal sebagai *instance* atau contoh, yang ditentukan oleh sebuah *tuple* (x,y), dimana x adalah himpunan atribut dan y adalah atribut tertentu, yang dinyatakan sebagai label kelas (juga dikenal sebagai kategori atau atribut target).
Beberapa teknik klasifikasi yang digunakan adalah *decision tree classifier*, *rule-based classifier*, *neural-network*, *support vector machine*, dan *naive bayes classifier*. Setiap teknik menggunakan algoritma pembelajaran untuk mengidentifikasikan model yang memberikan hubungan yang paling sesuai antara himpunan atribut dan label kelas dari data *input*.

2.4 Decision Tree

Decision tree (pohon keputusan) adalah sebuah diagram alir yang mirip dengan struktur pohon, di mana setiap internal *node* menotasikan atribut yang diuji, setiap cabangnya merepresentasikan hasil dari atribut tes tersebut, dan *leaf node* merepresentasikan kelas-kelas tertentu atau distribusi dari kelas-kelas (Han & Kamber, 2001).
Decision tree terdapat 3 jenis *node*, yaitu:

a. Root Node, merupakan *node* paling atas, pada *node* ini tidak ada *input* dan bisa tidak mempunyai *output* atau mempunyai *output* lebih dari satu.
b. Internal Node, merupakan *node* percabangan, pada *node* ini hanya terdapat satu *input* dan mempunyai *output* minimal dua.
c. Leaf node atau *terminal node* , merupakan *node* akhir, pada *node* ini hanya terdapat satu input dan tidak mempunyai *output*.

Penggunaan *decision tree* perlu memperhatikan hal-hal seperti, atribut mana yang akan dipilih untuk pemisahaan obyek, urutan atribut mana yang akan dipilih terlebih dahulu, struktur *tree*, *criteria* pemberhentian dan *pruning*.

2.5 Algoritma J48

Algoritma J48 adalah salah satu kelas dipaket *classifier* pada aplikasi *data mining* weka yang mengimplementasikan algoritma C4.5. Dalam membangun model berupa pohon keputusan Algoritma C4.5 menggunakan pendekatan teori *information gain*. Algoritm C4.5 mempunyai kelebihan karena dapat menghasilkan model berupa pohon. Model yang dihasilkan dengan Algoritma C4.5 (algoritma J48 dalam WEKA) yang dihasilkan dalam proses *training* dari data pelatihan berupa sebuah pohon keputusan. Pada algoritma C4.5, pemilihan atribut yang akan diproses menggunakan *information gain*. Jika dalam memilih atribut untuk memecah obyek dalam beberapa kelas harus kita pilih atribut yang menghasilkan *information gain* paling besar.

Ukuran *information gain* digunakan untuk memilih atribut uji pada setiap *node* di dalam *tree*. Ukuran ini digunakan untuk memilih atribut atau *node* pada pohon. Atribut dengan nilai *information gain* tertinggi akan terpilih sebagai parent bagi *node* selanjutnya. Sebelum menghitung *gain* harus dihitung terlebih dahulu nilai *entropy*-nya. *Entropy* adalah suatu parameter untuk mengukur *heterogenitas* (keberagaman) dari suatu kumpulan data sampel. Apabila sampel data semakin *heterogen* maka nilai dari *entropy*-nya semakin besar. Formula dari *entropy* adalah:

$$\text{Entropy } (S) = \sum_i^c -\text{pi} \log 2 \text{ pi}$$

dengan *c* adalah jumlah nilai yang ada pada atribut target (jumlah kelas klasifikasi), dan *pi* adalah jumlah sampel untuk kelas *i*.

Setelah nilai *entropy* diperoleh maka langkah selanjutnya adalah menghitung gain untuk mengukur efektifitas suatu atribut dalam mengklasifikasi data. Gain dihitung dengan menggunakan rumus:

$$Gain(S,A) = Entropy(s) - \sum_{v \in values(A)} \frac{Sv}{S} * Entropy(Sv)$$

Dimana *v* adalah suatu nilai yang mungkin untuk atribut *A*, values (*A*) menyatakan himpunan nilai-nilai yang mungkin untuk atribut *A*. *Sv* menyatakan jumlah sampel data untuk nilai *v*, dan *S* merupakan jumlah seluruh dari sampel data, sedangkan *entropy*(*Sv*) merupakan *entropy* untuk sampel-sampel yang memiliki nilai *v*.

2.6 Evaluasi dan Alat Ukur

Evaluasi model merupakan yang dikerjakan dalam penelitian ini dengan tujuan untuk memperoleh informasi yang terdapat pada hasil klasifikasi terhadap algoritma yang digunakan. Dalam *weka classifier* hasil klasifikasi yang diperoleh disertakan dengan beberapa alat ukur yang tersedia didalamnya, seperti *Confusion Matrix.*

Dalam penelitian ini dipilih alat ukur berupa *confusion matrix* yang terdapat pada *weka classifier* dengan tujuan untuk mempermudah dalam menganalisis performa algoritma karena *confusion matrix* memberikan informasi dalam bentuk angka sehingga dapat dihitung rasio keberhasilan klasifikasi. *Confusion matrix* adalah salah satu alat ukur berbentuk *matrix* 2x2 yang digunakan untuk mendapatkan jumlah ketepatan klasifikasi *dataset* terhadap kelas lulus dan tidak lulus pada algoritma yang dipakai. tiap kelas yang diprediksi memiliki empat kemungkinan keluaran yang berbeda, yaitu *trues positives* (TP) dan *true negatives* (TN) yang menunjukkan ketepatan klasifikasi. Jika prediksi keluaran bernilai positif sedangkan nilai aslinya adalah negatif maka disebut dengan *false positive* (FP) dan jika prediksi keluaran bernilai negatif sedangkan nilai aslinya adalah positif maka disebut dengan *false negative* (FN). Berikut ini pada tabel 2.1 disajikan bentuk *confusion matrix* seperti yang telah dijelaskan sebelumnya.

Tabel 1. Confusion matrix untuk masalah klasifikasi kelas

		Predicted Class	
Actual Class		Yes	No
	Yes	True Positive	False Negative
	No	False Positive	True Negative

Beberapa kegiatan yang dapat dilakukan dengan menggunakan data hasil klasifikasi dalam *confusion matrix* diantaranya:

- Menghitung nilai rata-rata keberhasilan klasifikasi (*overall success rate*) ke dalam kelas yang sesuai dengan cara membagi jumlah data yang terklasifikasi dengan benar, dengan seluruh data yang diklasifikasi.
- Selain itu dilakukan pula perhitungan persentase kelas positif (*true positive dan false positive*) yang

diperoleh dalam klasifikasi, yang disebut dengan *lift chart*

- *Lift chart* terkait erat dengan sebuah teknik dalm mengevaluasi skema data mining yang dikenal dengan ROC (*receiver operating characteristic*) yang berfungsi mengekspresikan persentase jumlah proporsi positif dan negatif yang diperoleh.
- *Recall precision* berfungsi menghitung persentase *false positive* dan *false negative* untuk menemukan informasi di dalamnya

3. METODE PENELITIAN

Metode yang digunakan dalam penelitian ini adalah sebagai berikut:

a. Pengamatan (Observasi)
Melakukan pengamatan langsung ke bagian UPT – SIM pengolahan data Universitas Bina Darma Palembang untuk mendapatkan data yang dibutuhkan.

b. Wawancara (*Interview*)
Mengadakan wawancara dengan pihak-pihak yang berkaitan langsung dengan permasalahan yang sedang di bahas pada tugas akhir ini untuk memperoleh gambaran dan penjelasan secara mendasar.

c. Studi Pustaka
Penulis mengumpulkan berbagai referensi dan literatur pendukung penelitian berupa buku, jurnal dan artikel yang berasal dari berbagai sumber yang erat kaitannya dengan objek permasalahan

Tahapan data mining yang digunakan dalam penelitian ini meliputi: proses seleksi, pembersihan data, integrasi data, transformasi data, data mining dan evaluasi pola serta presentasi pengetahuan.

3.1 Seleksi Data (data selection)

Sumber data yang digunakan dalam penelitian ini berasal dari data mahasiswa tahun 2000 sampai dengan 2011 pada jurusan sistem informasi Universitas Bina Darma Palembang. Beberapa jenis data diperoleh dari sistem yang berjalan namun hanya data mahasiswa dan data IPK (Indeks Prestasi Komulatif) saja yang digunakan untuk penelitian, karena informasi yang terkandung didalamnya sudah mewakili informasi yang dibutuhkan untuk dijadikan *indicator* penentu klasifikasi data keluaran yang diinginkan.
Jumlah data yang diperoleh adalah sebanyak 1.000 *record* data yang berasal dari tabel mahasiswa dan 41.367 record data yang berasal dari tabel IPK. Dataset mahasiswa terdiri dari 23 atribut yang menjelaskan identitas diri mahasiswa dan informasi tentang keadaan mahasiswa yang bersangkutan. *Dataset* mahasiswa diambil dari penggabungan beberapa tabel yang terdiri dari tb_mhs, tb_khs, dan tb_mk. Atribut tersebut diantaranya adalah nim, nama, jenis_kelamin, temp_lahir, tgl_lahir, kd_progdi, alamat, asal_sek, kota, pek_ortu, kd_mk, sms, sms_pendek, tahun_akademik, kelas, kd_dosen, kd_progdi, tugas, kuis, mid, semester, nilai_angka, nilai_huruf, dan sks. Sedangkan dataset IPK hanya terdiri dari 5atribut, dimana memberikan informasi mengenai prestasi akademik dan beban studi yang diambil mahasiswa yang bersangkutan. Atribut tersebut diantaranya adalah nim, nama, sum(sks), total_nilai, dan IPK.

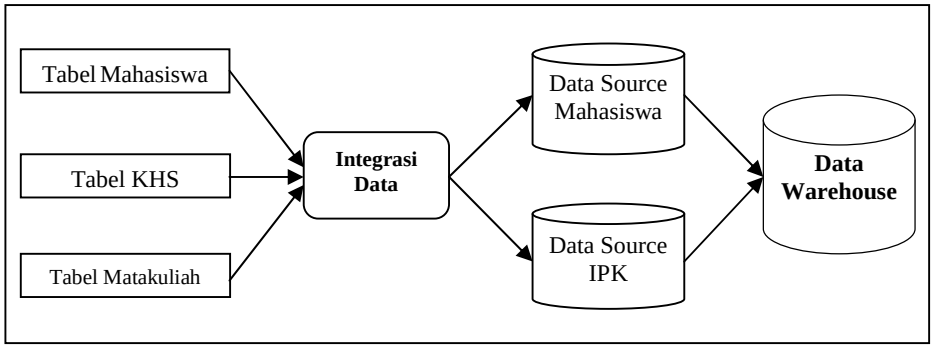
3.2 Praproses Data

a. Pembersihan Data (*data cleaning*)

Tahap kedua pada proses data mining adalah *cleaning* data yaitu melakukan pembersihan data terhadap *noise* yang ditemukan berupa *missing value*, inkonsisten data, dan *redundant* data. Seluruh atribut diatas selanjutnya akan diseleksi untuk mendapatkan atribut-atribut yang berisi nilai yang relevan, tidak *missing value*, dan tidak *redundant*, dimana ketiga syarat tersebut merupakan syarat awal yang harus dikerjakan dalam *data mining* sehingga akan diperoleh *dataset* yang bersih untuk digunakan pada tahap *mining* data. Pada *dataset* mahasiswa ditemukan data yang *missing value* maka dilakukan *cleaning data* terhadap data dengan *missing value* yang dimaksud.

b. Integrasi Data

Tahap ketiga pada proses *data mining* adalah tahap integrasi data yaitu tahap penggabungan data dengan tujuan memindahkan seluruh data yang telah di-*cleaning* menjadi satu tabel. Pada tahap ini dari ketiga tabel akan di integrasikan untuk mendapatkan data *source* mahasiswa dan data *source* IPK, dari kedua data *source* tersebut akan digabungkan dalam satu tabel yang nantinya akan memudahkan dalam proses mining data. Gambar 1 menunjukkan bentuk integrasi data dari tabel mahasiswa, tabel khs, dan tabel matakuliah yang membentuk satu tabel tunggal sebagai data warehouse yang akan digunakan untuk proses analisis teknik data mining.



Gambar 1 Proses Integrasi Data

c. Transformasi Data (*Tranformation*)
Tahap keempat pada proses data mining adalah tahap tranformasi data yaitu pada tahap ini data diubah menjadi bentuk yang sesuai untuk diproses dalam *data mining*. Dalam penelitian ini data yang akan diproses dari *database* mysql akan diubah menjadi *file* CSV(*Comma Separated Values*) yang dapat digunakan untuk pengolahan data pada *tools* WEKA.

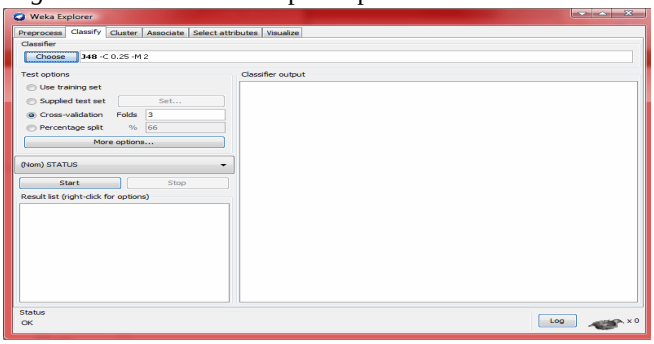
4. HASIL DAN PEMBAHASAN

4. 1 *Data Mining*

Data mining merupakan proses mencari pola atau informasi menarik dalam data yang terpilih dengan menggunakan teknik atau metode tertentu. Pemilihan teknik dan algoritma yang tepat sangat bergantung pada proses KDD (*Knowledge Discovery in Database*) secara keseluruhan. Pada penelitian ini penerapan data mining menggunakan teknik *classification* dan algoritma J48 untuk mengetahui vareiabel penentu tingkat kelulusan mahasiswa. Tahap *data mining* merupakan inti dari tahapan KDD yang dilakukan untuk menganalisis data yang sudah bersih. Proses análisis *data mining* dalam penelitian ini menggunakan perangkat lunak aplikasi *data mining* WEKA 3.6.8. Dengan menggunakan *data warehouse* yang telah dihasilkan melalui tahapan awal dalam *data mining* maka langkah berikutnya adalah melakukan proses análisis *data mining* dengan teknik *classification* dan menggunakan algoritma J48.

4.2 Penerapan Algoritma J48 menggunakan Aplikasi WEKA

Berikut ini merupakan tampilan proses klasifikasi J48 dengan menggunakan atribut dalam *data warehouse* yang sudah melalui proses *cleaning* dan transformasi data pada aplikasi WEKA 3.6.8.



Gambar 2. Tampilan Algoritma J48 dalam aplikasi WEKA

Pada gambar diatas beberapa fungsi atribut yang terdapat dalam aplikasi WEKA adalah sebagai berikut:
Keterangan :

- *Choose* : J48-C0.25-M2 maksudnya adalah pemilihan algoritma klasifikasi
- *Use Training Set* : Menggunakan *data training set*
- *Supplied test set* : Menggunakan *data testing*
- *Cross-validation* : Membagi data menurut bagian
- *Percentage Split* : Persentase dari perpecahan atau percabangan

Dengan menggunakan *test options 3-fold cross validation* maka *dataset* tersebut kemudian digunakan untuk mengkontruksi pohon keputusan (*decisión tree*) yang dimulai dengan pembentukan bagian akar, kemudian data terbagi berdasarkan atribut-atribut yang sesuai untuk dijadikan *leaf node*. Tahap ini dimulai dengan melakukan

seleksi atribut menggunakan formula *information gain* yang terdapat pada algoritma J48, sehingga diperlukan nilai *gain* untuk masing-masing atribut, yang mana atribut dengan nilai *gain* tertinggi akan menjadi *parent* bagi *node-node* selanjutnya. *Node-node* tersebut berasal dari atribut-atribut yang memiliki nilai *gain* yang lebih kecil dari nilai *gain* atribut *parent*. Maka untuk mendapatkan nilai *gain* dari dua kelas *output* yang berbeda yaitu ***lulus*** dan ***tidak lulus*** pada *dataset* mahasiswa adalah dengan menghitung tingkat *impurity* kedua kelas tersebut. Berikut ini adalah contoh perhitungan *node* data kelas mahasiswa ***lulus*** dan ***tidak lulus*** berdasarkan beberapa atribut.

Tabel 2. Perhitungan Node

Node			Jumlah Data	Lulus	Tidak Lulus	Entropy	Gain
	Total		774	557	217	0,86	
	PekOrtu						0,07
		Buruh	50	45	5	0,47	
		Pensiun	28	26	2	0,37	
		Petani	42	38	4	0,45	
		Pns	287	210	77	0,83	
		Swasta	193	149	44	0,77	
		Wiraswasta	174	109	65	0,95	
	Kota						0,02
		Palembang	596	404	192	0,92	
		Luar Kota	178	153	25	0,59	
	JenisKelamin						0,04
		Pria	501	321	180	0,94	
		Wanita	273	233	40	0,6	
	TempLahir						0,08
		Palembang	430	350	80	0,69	
		Luar Kota	344	238	106	0,89	
	AsalSek						0,06
		SMA	620	465	155	0,81	
		SMK	123	99	24	0,71	
		MAN	31	17	14	0,99	

Berikut ini adalah contoh perhitungan untuk mendapatkan nilai entropy dan gain yang diberikan oleh tabel, dengan mengambil contoh atribut jenis kelamin berdasarkan rumus yang ada maka dapat dihitung nilai entropy dan gain-nya sebagai berikut:

$$\text{Entropy } (S) = \sum_{i}^{c} -\text{pi log2 pi}$$

Entropy (S)=$(-\frac{557}{774} * log2\left(\frac{557}{774}\right))+(-\frac{217}{774} * log2\left(\frac{217}{774}\right))$

Entropy (S) = 0,86

Sementara itu, nilai *Gain* pada baris jenis kelamin dihitung dengan menggunakan formula *gain* sebagai berikut:

$$Gain(S,A)=Entropy(s) - \sum_{v \in values(A)} \frac{Sv}{S} * Entropy(Sv)$$

Gain(S,JenisKelamin)=0,86 –$((\frac{501}{774} * (0,94) + (\frac{273}{774} * (0,60))$

Gain(S,JenisKelamin)=0,04)

Hasil diatas diperoleh dengan menggunakan *test options 3-fold cross validation*. Dari hasil tabel tersebut diperoleh atribut dengan nilai *gain* tertinggi yang kemudian dipilih sebagai simpul pertama pada *decision tree*. Pada simpul selanjutnya secara berurutan diisi oleh atribut-atribut yang bernilai *gain* lebih rendah, dan akan berhenti pada simpul akhir yang berisi kelas *output* dari setiap cabangnya yang dikenal dengan nama *leaf* atau daun. Tabel 2 diatas tersebut menyajikan nilai *gain* dari seluruh atribut yang mana nilai *gain* atribut pek_ortu, kota, jenis_kelamin, temp_lahir dan asal_sek yang terdapat dalam tabel adalah hasil pembulatan terhadap nilai aslinya.

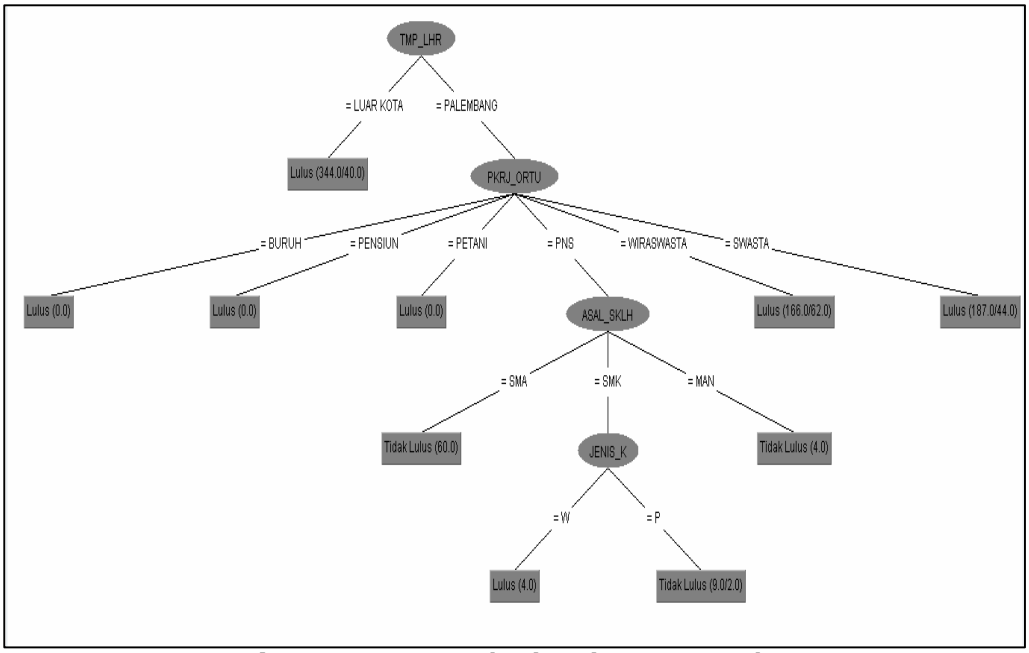
Pada tabel 2 terlihat bahwa nilai atribut temp_lahir memiliki nilai *gain* tertinggi, sehingga atribut ini menjadi atribut *root* pada *decision tree*, kemudian dilanjutkan dengan atribut pek_ortu, asal_sek dan jenis kelamin dan diakhiri oleh label kelas lulus dan tidak lulus yang berfungsi sebagai *leaf*. Maka dapat dikatakan bahwa parameter penentu pertama seorang mahasiswa tingkat kelulusan pada waktu yang akan datang dilihat dari tempat lahir mahasiswa yang bersangkutan, kemudian pekerjaan orang tua, asal sekolah, dan jenis kelamin mahasiswa tersebut. Sedangkan atribut pada kota nilai *gain* yang diperoleh sangat kecil jika dibandingkan

dengan atribut lainnya, sehingga dapat disimpulkan bahwa dukungan informasi yang terkandung dalam atribut tersebut terhadap *output* yang dicapai sangat kecil. Maka atribut tersebut akan dipangkas (*pruned*) dan tidak terpilih sebagai atribut untuk *decision tree*. Pada gambar 3 berikut ini merupakan hasil klasifikasi data menggunakan *weka classifier*.

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     ASLI_weka2-1-weka.filters.unsupervised.attribute.Remove-R1
Instances:    774
Attributes:   6
              JENIS_K
              TMP_LHR
              ASAL_SKLH
              KOTA
              PKRJ_ORTU
              STATUS
Test mode:3-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

TMP_LHR = LUAR KOTA: Lulus (344.0/40.0)
TMP_LHR = PALEMBANG
|   PKRJ_ORTU = BURUH: Lulus (0.0)
|   PKRJ_ORTU = PENSIUN: Lulus (0.0)
|   PKRJ_ORTU = PETANI: Lulus (0.0)
|   PKRJ_ORTU = PNS
|   |   ASAL_SKLH = SMA: Tidak Lulus (60.0)
|   |   ASAL_SKLH = SMK
|   |   |   JENIS_K = W: Lulus (4.0)
|   |   |   JENIS_K = P: Tidak Lulus (9.0/2.0)
|   |   ASAL_SKLH = MAN: Tidak Lulus (4.0)
|   PKRJ_ORTU = WIRASWASTA: Lulus (166.0/62.0)
|   PKRJ_ORTU = SWASTA: Lulus (187.0/44.0)
```

Gambar 3. Hasil Klasifikasi menggunakan WEKA

Pada gambar 3 diatas terlihat bahwa *weka classifier* hanya memilih atribut temp_lahir, pek_ortu, asal_sek, dan jenis_kelamin sebagai atribut dalam *decision tree*, sedangkan atribut pada kota langsung terpangkas dari *decision tree*. Maka dapat disimpulkan bahwa dengan jumlah dan jenis data yang ada hanya dibutuhkan beberapa atribut untuk mendapatkan kelas output dari dataset tersebut. *Decision tree* yang dihasilkan, seperti gambar 4.



Gambar 4. Decision Tree hasil Analisis Menggunakan WEKA

5. KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan yang telah dijelaskan maka dapat disimpulkan bahwa atribut tempat lahir adalah atribut yang memiliki nilai *Gain* tertinggi, sehingga atribut ini menjadi *root* pada *Decision Tree*. Variabel tempat lahir, pekerjaan orang tua, asal sekolah dan jenis kelamin adalah variabel yang menentukan tinggkat kelulusan mahasiswa pada waktu yang akan datang. Rata-rata klasifikasi algoritma J48 dalam melakukan klasifikasi data mencapai akurasi diatas 90%. Hal ini menunjukkan bahwa algoritma tersebut memiliki performa yang handal dalam melakukan klasifikasi.

DAFTAR PUSTAKA

Angga Raditya, *Implementasi Data Mining Classification untuk Menacrai Pola Prediksi Hujan dengan Menggunakan Algoritma C4.5*, Jurusan Teknik Informatika, Fakultas Teknologi Industri, Universitas Gunadarma

Ariana Azimah, Yudho Giri Sucahyo, *Penggunaan Data Warehouse Dan Data Mining Untuk Data Akademik Sebuah Studi Kasus Pada Universitas Nasional*, 2007

Daniel T, Larose, 2005. "*Discovering Knowledge in Data: An Introduction to Data Mining"*. John Wiley & Sons, Inc

Han, J. And Kamber, M, 2006, "*Data Mining Concept and Techniques Second Edition "*. Morgan Kauffman, San Francisco.

Iin Ernawati, 2008, "*Prediksi Status Keaktifan Studi Mahasiswa dengan Algoritma C5.0 dan K-Nearest Neighbor*", Institut Pertanian Bogor

Rainardi, Vincent, 2008, "*Building Data Warehouse with Examples in SQL Server*", Springer, New York.

Tan S, Kumar P, Steinbach M. 2005. "*Introduction To Data Mining"*. Addison Wesley